

## DEVIATION OPTIMAL LEARNING USING GREEDY $Q$ -AGGREGATION

BY DONG DAI, PHILIPPE RIGOLLET<sup>1</sup> AND  
TONG ZHANG

*Rutgers University, Princeton University and Rutgers University*

Given a finite family of functions, the goal of model selection aggregation is to construct a procedure that mimics the function from this family that is the closest to an unknown regression function. More precisely, we consider a general regression model with fixed design and measure the distance between functions by the mean squared error at the design points. While procedures based on exponential weights are known to solve the problem of model selection aggregation in expectation, they are, surprisingly, sub-optimal in deviation. We propose a new formulation called  $Q$ -aggregation that addresses this limitation; namely, its solution leads to sharp oracle inequalities that are optimal in a minimax sense. Moreover, based on the new formulation, we design greedy  $Q$ -aggregation procedures that produce sparse aggregation models achieving the optimal rate. The convergence and performance of these greedy procedures are illustrated and compared with other standard methods on simulated examples.

**1. Introduction.** Model selection is one of the major aspects of statistical learning and, as such, has received considerable attention over the past decades. More recently, the seminal works of Nemirovski (2000) and Tsybakov (2003) have introduced an idealized setup to study the properties of model selection procedures independently of the models themselves. We consider this so-called *pure model selection aggregation* (or simply MS aggregation) framework for the simple model of Gaussian regression with fixed design.

Let  $x_1, \dots, x_n$  be  $n$  given design points in a space  $\mathcal{X}$ , and let  $\mathcal{H} = \{f_1, \dots, f_M\}$  be a given dictionary of real valued functions on  $\mathcal{X}$ . The goal is to

---

Received March 2012; revised June 2012.

<sup>1</sup>Supported in part by NSF Grants DMS-09-06424 and CAREER-DMS-1053987.

*AMS 2000 subject classifications.* Primary 62G08; secondary 90C52, 62G05, 62G20.

*Key words and phrases.* Regression, model selection, model averaging, greedy algorithm, exponential weights, oracle inequalities, deviation bounds, lower bounds, deviation suboptimality.

This is an electronic reprint of the original article published by the  
Institute of Mathematical Statistics in *The Annals of Statistics*,  
2012, Vol. 40, No. 3, 1878–1905. This reprint differs from the original in  
pagination and typographic detail.

estimate an unknown regression function  $\eta: \mathcal{X} \rightarrow \mathbb{R}$  at the design points based on observations

$$Y_i = \eta(x_i) + \xi_i,$$

where  $\xi_1, \dots, \xi_n$  are i.i.d.  $\mathcal{N}(0, \sigma^2)$ . Our main results are actually stated for sub-Gaussian random variables, but since most of the literature is available only for Gaussian noise, we temporarily make this assumption to ease comparisons throughout the [Introduction](#). The performance of an estimator  $\hat{\eta}$  is measured by its mean square error (MSE) defined by

$$\text{MSE}(\hat{\eta}) = \frac{1}{n} \sum_{i=1}^n (\eta(x_i) - \hat{\eta}(x_i))^2.$$

In the pure model selection aggregation framework, the goal is to build an estimator  $\hat{\eta}$  that mimics the function  $f_j$  in the dictionary with the smallest MSE. Formally, a good estimator  $\hat{\eta}$  should satisfy the following *oracle inequality* in a certain probabilistic sense:

$$(1.1) \quad \text{MSE}(\hat{\eta}) \leq \min_{j=1, \dots, M} \text{MSE}(f_j) + \Delta_{n,M}(\sigma^2),$$

where the remainder term  $\Delta_{n,M} > 0$  should be as small as possible. Note that oracle inequality (1.1) is a truly finite sample result, and the remainder term should show the interplay between the three fundamental parameters of the problem: the “dimension”  $M$ , the sample size  $n$  and the noise level  $\sigma^2$ . Most oracle inequalities for model selection aggregation have been produced in expectation [see the references in Rigollet and Tsybakov (2012)] with notable exceptions [Audibert (2008), Lecué and Mendelson (2009), Gaïffas and Lecué (2011), Dai and Zhang (2011), Rigollet (2012)] who produced oracle inequalities that hold with high probability and to which we will come back later.

From the early days of the pure model selection problem, it has been established [see, e.g., Tsybakov (2003), Rigollet (2012)] that the smallest possible order for  $\Delta_{n,M}(\sigma^2)$  was  $\sigma^2(\log M)/n$  for oracle inequalities in expectation, where “smallest possible” is understood in the following minimax sense. There exists a dictionary  $\mathcal{H} = \{f_1, \dots, f_M\}$  such that the following holds. For any estimator  $\hat{\eta}$ , there exists a regression function  $\eta$  such that

$$\mathbb{E} \text{MSE}(\hat{\eta}) \geq \min_{j=1, \dots, M} \text{MSE}(f_j) + C\sigma^2 \frac{\log M}{n}$$

for some positive constant  $C$ . Moreover, it follows from the same results that this lower bound holds not only in expectation but also with positive probability.

The established terminology *model selection* is somewhat misleading. Indeed, while the goal is to mimic the best model in the dictionary  $\mathcal{H}$ , it

has been shown [see [Rigollet and Tsybakov \(2012\)](#), Theorem 2.1] that there exists a dictionary  $\mathcal{H}$  such that any estimator (selector)  $\hat{\eta}$  restricted to be one of the elements of  $\mathcal{H}$  cannot satisfy an oracle inequality such as (1.1) with a remainder term of order smaller than  $\sigma\sqrt{(\log M)/n}$ , which is clearly suboptimal. Rather than model selection, *model averaging* has been successfully employed to derive oracle inequalities in expectation such as (1.1). More precisely, model averaging consists in choosing  $\hat{\eta}$  as a convex combination of the  $f_j$ s with carefully chosen weights. Let  $\Lambda$  be the flat simplex of  $\mathbb{R}^M$  defined by

$$\Lambda = \left\{ \lambda = (\lambda_1, \dots, \lambda_M)^\top \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1 \right\}.$$

Each  $\lambda \in \Lambda$  yields an *aggregate* estimator  $\hat{\eta} = f_\lambda$ , where

$$f_\lambda = \sum_{j=1}^M \lambda_j f_j.$$

This is why we refer to this problem as *model selection aggregation* or MS aggregation. The early papers of Catoni ([1999](#)) and Yang ([1999](#)) introduced and proved optimal theoretical guarantees for a model averaging estimator called *progressive mixture* that was later studied in Audibert ([2008](#)) and Juditsky, Rigollet and Tsybakov ([2008](#)) from various perspectives. This estimator is based on *exponential weights*, which, since then, have been predominantly used and have led to optimal oracle inequalities in expectation. Let  $\pi = (\pi_1, \dots, \pi_M)^\top \in \Lambda$  be a given *prior* and  $\beta > 0$  be a temperature parameter, then the  $j$ th exponential weight is given by

$$(1.2) \quad \lambda_j^{\text{EXP}} \propto \pi_j \exp(-n \widehat{\text{MSE}}(f_j)/\beta),$$

where

$$\widehat{\text{MSE}}(f_j) = \frac{1}{n} \sum_{i=1}^n (Y_i - f_j(x_i))^2.$$

The most common prior choice is the uniform prior  $\pi = (1/M, \dots, 1/M)^\top$ , but other choices that put more or less weight on different functions of the dictionary have been successfully applied to various related problems; see, for example, Dalalyan and Salmon ([2011](#)), [Rigollet and Tsybakov \(2011, 2012\)](#). Note that progressive mixture contains an extra averaging step which is irrelevant to the fixed design problem that we study here, but we implement it in Section 5 for comparison with the nonaveraged procedure.

The fixed design Gaussian regression was considered in Dalalyan and Tsybakov ([2007, 2008](#)) who proved an oracle inequality of the form (1.1)

with optimal remainder term. This result suffers two deficiencies: first, it can be extended to other types of noise, but not to sub-Gaussian distributions in general. Second, and perhaps most importantly, it holds only in expectation and not with high probability. While this second limitation may have followed the proof technique, we actually show in Section 3 that it is inherent to exponential weights. Consequently, we say that exponential weights are *deviation suboptimal* since the expectation of the resulting MSE is of the optimal order, but the deviations around the expectation are not. Note also that the original paper of Dalalyan and Tsybakov (2007) made some boundedness assumption on the distance between function in the dictionary  $\mathcal{H}$  and the regression function  $\eta$ . This assumption was lifted in their subsequent paper [Dalalyan and Tsybakov (2008)]. In this paper, we make no such assumption except for the lower bound, which, of course, makes our result even stronger.

For regression with random design, Audibert (2008) observed also that various progressive mixture rules are deviation suboptimal. In the same paper, he addressed this issue by proposing the STAR algorithm which is optimal both in expectation and in deviations under the uniform prior and, remarkably, does not require any parameter tuning as opposed to progressive mixture rules. Also for random design, Lecué and Mendelson (2009) followed by Gaïffas and Lecué (2011) proposed deviation optimal methods based on the same sample splitting idea. However, sample splitting method do not carry to fixed design. Subsequently, Rigollet (2012) proposed a new estimator, similar to the one studied in the rest of the paper and that enjoys the same theoretical properties as the STAR algorithm but for fixed design regression. However, while it is the solution of a convex optimization problem, Rigollet's method comes without implementation. Finally, a first implementation of a greedy algorithm that enjoys optimal deviation was proposed in Dai and Zhang (2011). Our subsequent results extend both the results of Rigollet (2012) and Dai and Zhang (2011) in various directions.

In Section 2 of the present paper, we study the deviation suboptimality of two commonly used aggregate estimators: the aggregate by exponential weights and the aggregate by projection. Then, in Section 3, we extend the original method of Rigollet (2012) in several directions. First and foremost, our extension allows us to put a prior weight on each element of the dictionary. These prior weights appear explicitly in the oracle inequalities that are derived in Section 3. Both the method of Rigollet (2012) and ours are solutions of convex optimization problems. In Section 4, we propose efficient greedy model averaging (GMA) procedures that approximately solve the newly proposed  $Q$ -aggregation formulations. It is shown that GMA can produce sparse model aggregates that achieve optimal deviation bounds. The performance of different model selection and aggregation estimators are compared in Section 5.

NOTATION. For any vector  $v$ , we denote by  $v_j$  its  $j$ th coordinate. Moreover, for any functions  $f, g: \mathcal{X} \rightarrow \mathbb{R}$ , we define the pseudo-norm

$$\|f\|^2 = \frac{1}{n} \sum_{i=1}^n f(x_i)^2,$$

and the associated inner product

$$\langle f, g \rangle = \frac{1}{n} \sum_{i=1}^n f(x_i)g(x_i).$$

Also, we define the function  $\mathbf{Y}: \mathcal{X} \rightarrow \mathbb{R}$  to be any function such that  $\mathbf{Y}(x_i) = Y_i$ . Observe that with the above notation, we have

$$\widehat{\text{MSE}}(f) = \|\mathbf{Y} - f\|^2, \quad \text{MSE}(f) = \|\eta - f\|^2.$$

Finally, for any  $p \geq 1$ , we denote by  $|\cdot|_p$  the  $\ell_p$  norm.

**2. Deviation suboptimality of commonly used estimators.** It is well known [see, e.g., Rigollet and Tsybakov (2012)] that the exponential weights  $\lambda^{\text{EXP}}$  defined in (1.2) are the solution of the following minimization problem:

$$(2.1) \quad \lambda^{\text{EXP}} \in \operatorname{argmin}_{\lambda \in \Lambda} \left\{ \sum_{j=1}^M \lambda_j \widehat{\text{MSE}}(f_j) + \frac{\beta}{n} \sum_{j=1}^M \lambda_j \log\left(\frac{\lambda_j}{\pi_j}\right) \right\}.$$

It was shown in Dalalyan and Tsybakov (2007, 2008) that for  $\beta \geq 4\sigma^2$ , it holds

$$(2.2) \quad \mathbb{E} \text{MSE}(f_{\lambda^{\text{EXP}}}) \leq \min_{j=1, \dots, M} \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log(\pi_j^{-1}) \right\}.$$

The proof of this result relies heavily on the fact that the oracle inequality holds in expectation and whether the result also holds with high probability arises as a natural question. While the paper of Audibert (2008) does not cover the fixed design Gaussian regression framework of our paper and concerns exponential weights with an extra averaging step, it contributed to the common belief that exponential weights would be suboptimal in deviation. In particular, Lecué and Mendelson (2012) derived lower bounds for the performance of exponential weights in expectation when  $\beta$  is chosen below a certain *constant* threshold in the case of regression with random design. Moreover, they proved deviation sub-optimality of exponential weights when  $\beta$  is less than  $\sqrt{n}/(\log n)$ . However, these lower bounds rely heavily on the fact that the design is random and do not extend to the fixed design case. In particular, their construction uses  $Y \equiv 0$ , which is clearly an easy problem in the fixed design case. Proposition 2.1 states precisely that exponential weights are deviation suboptimal, if  $\beta$  is chosen small enough and in particular if  $\beta$  is *any* constant with respect to  $M$  and  $n$ .

Another natural solution to solve the MS aggregation problem is to take the vector of weights  $\lambda^{\text{PROJ}}$  defined by

$$(2.3) \quad \lambda^{\text{PROJ}} \in \underset{\lambda \in \Lambda}{\operatorname{argmin}} \widehat{\text{MSE}}(\mathbf{f}_\lambda).$$

We call  $\lambda^{\text{PROJ}}$  the vector of *projection weights* since the aggregate estimator  $\mathbf{f}_{\lambda^{\text{PROJ}}}$  is the projection of  $\mathbf{Y}$  onto the convex hull of the  $f_j$ s.

It has been established that this choice is *near-optimal* for the more difficult problem of *convex aggregation* with fixed design [see Juditsky and Nemirovski (2000), Nemirovski (2000), Rigollet (2012)] where the goal is to mimic the best convex combination of the  $f_j$ s as opposed to simply mimicking the best of them. More precisely, it follows from Theorem 3.5 in Rigollet (2012) that

$$\begin{aligned} \mathbb{E} \text{MSE}(\mathbf{f}_{\lambda^{\text{PROJ}}}) &\leq \min_{\lambda \in \Lambda} \text{MSE}(\mathbf{f}_\lambda) + 2\sigma \sqrt{\frac{\log M}{n}} \\ &\leq \min_{j=1, \dots, M} \text{MSE}(f_j) + 2\sigma \sqrt{\frac{\log M}{n}}, \end{aligned}$$

and a similar oracle inequality also holds with high probability. The second inequality is very coarse, and it is therefore natural to study whether a finer analysis of this estimator would yield an optimal oracle inequality for the aggregate  $\mathbf{f}_{\lambda^{\text{PROJ}}}$  both in expectation and with high probability. This question was investigated by Lecué and Mendelson (2009) who proved that  $\mathbf{f}_{\lambda^{\text{PROJ}}}$  cannot satisfy an oracle inequality of the form (1.1) with high probability and with a remainder term  $\Delta_{n,M}(\sigma^2)$  of order smaller than  $n^{-1/2}$ . Their proof, however, heavily uses the fact that the design is random, and we extend it to the fixed design case in Proposition 2.2 below.

For both aggregates considered below, we use the following notation. For each  $j = 1, \dots, M$ , we identify the functions  $f_j$  on  $\{x_1, \dots, x_n\}$  with a vector  $\mathbf{f}_j = (f_j(x_1), \dots, f_j(x_n)) \in \mathbb{R}^n$  where we systematically use the gothic font to identify such vectors throughout the rest of the section. Moreover, for any vector of weights  $\lambda \in \mathbb{R}^M$ , we write  $\mathbf{f}_\lambda = (\mathbf{f}_\lambda(x_1), \dots, \mathbf{f}_\lambda(x_n))$ .

**2.1. Aggregate by exponential weights.** Consider the following dictionary  $\mathcal{H}$ . Assume that  $M, n \geq 3$ . Let  $\mathbf{e}^{(1)} = (1, 0, \dots, 0)^\top \in \mathbb{R}^n$  and  $\mathbf{e}^{(2)} = (0, 1, 0, \dots, 0)^\top \in \mathbb{R}^n$  be the first two vectors of the canonical basis of  $\mathbb{R}^n$ . Moreover, let  $\mathbf{e}^{(3)}, \dots, \mathbf{e}^{(M)} \in \mathbb{R}^n$  be  $M - 2$  unit vectors of  $\mathbb{R}^n$  that are orthogonal to both  $\mathbf{e}^{(1)}$  and  $\mathbf{e}^{(2)}$ . Let  $\mathbf{f}_1, \dots, \mathbf{f}_M$  be such that

$$\mathbf{f}_1 = \sigma \sqrt{n} \mathbf{e}^{(1)}, \quad \mathbf{f}_2 = \sigma(1 + \sqrt{n}) \mathbf{e}^{(2)},$$

and for any  $j = 3, \dots, M$ ,  $\mathbf{f}_j$  is defined by

$$\mathbf{f}_j = \mathbf{f}_2 + \sigma \alpha_j \mathbf{e}^{(j)},$$

where  $\alpha_3, \dots, \alpha_M \geq 0$  are tuning parameters to be chosen later. Moreover, take the regression function  $\eta \equiv 0$  so that  $\text{MSE}(f_1) \leq \text{MSE}(f_j)$  for any  $j \geq 2$ . Observe that  $\|f_j\| \geq \sigma$  so that the following lower bounds cannot be interpreted as artifacts of scaling the signal-to-noise ratio.

Assume that  $M \geq 4$  and  $n \geq 3$ . We call low temperatures, parameters  $\beta > 0$  such that

$$(2.4) \quad \beta \leq \frac{2\sigma^2\sqrt{n}}{\log(8\sqrt{n})}.$$

In particular the exponential weights employed in the literature on MS aggregation use the low temperature  $\beta = 4\sigma^2$ ; see, for example, (2.2) above.

**PROPOSITION 2.1.** *Fix  $M \geq 4, n \geq 3$  and assume that the noise random variables  $\xi_1, \dots, \xi_n$  are i.i.d.  $\mathcal{N}(0, \sigma^2)$ . Let  $\eta$  and  $\mathcal{H}$  be defined as above. Then, the aggregate estimator  $\mathbf{f}_{\lambda^{\text{EXP}}}$  with exponential weights  $\lambda^{\text{EXP}}$  given by (1.2) satisfies*

$$\text{MSE}(\mathbf{f}_{\lambda^{\text{EXP}}}) \geq \min_{j=1, \dots, M} \text{MSE}(f_j) + \frac{\sigma^2}{4\sqrt{n}},$$

with probability at least 0.07 at low temperatures, for any  $\alpha_3, \dots, \alpha_M \geq 0$ .

Moreover, if  $M \geq 8\sqrt{n}$  and for any  $j \geq 3$ , we have

$$(2.5) \quad 2\sqrt{2\log(100M)} \leq \alpha_j \leq n^{1/4},$$

then, the same result holds at any temperature, with probability at least 0.06.

**PROOF.** Note first that by homogeneity, one may assume that  $\sigma = 1$ . Moreover, write for simplicity  $\lambda = \lambda^{\text{EXP}}$ . If we assume  $\lambda_1 \leq 1/2$ , we obtain

$$(2.6) \quad \begin{aligned} |\mathbf{f}_\lambda|_2^2 - |\mathbf{f}_1|_2^2 &\geq |\lambda_1 \mathbf{f}_1 + (1 - \lambda_1) \mathbf{f}_2|_2^2 - |\mathbf{f}_1|_2^2 \\ &= (1 - \lambda_1)^2 |\mathbf{f}_2|_2^2 - (1 - \lambda_1^2) |\mathbf{f}_1|_2^2 \\ &\geq 2(1 - \lambda_1)^2 \sqrt{n} + [(1 - \lambda_1)^2 - (1 - \lambda_1^2)]n \\ &\geq \sqrt{n}/2 - 2\lambda_1 n. \end{aligned}$$

We first treat the low temperature case where  $\beta$  is chosen as in (2.4). Define the event

$$E = \{n\widehat{\text{MSE}}(f_2) + 2\sqrt{n} \leq n\widehat{\text{MSE}}(f_1)\},$$

and observe that  $\eta \equiv 0$  gives

$$(2.7) \quad E = \{2\langle \mathbf{f}_2 - \mathbf{f}_1, \xi \rangle_2 \geq |\mathbf{f}_2|_2^2 - |\mathbf{f}_1|_2^2 + 2\sqrt{n}\}.$$

On the one hand, we have  $|\mathbf{f}_2|_2^2 - |\mathbf{f}_1|_2^2 = 1 + 2\sqrt{n}$ , and on the other hand

$$|\mathbf{f}_2 - \mathbf{f}_1|_2^2 = |\mathbf{f}_2|_2^2 + |\mathbf{f}_1|_2^2 = (2n + 2\sqrt{n} + 1) \geq \frac{1}{8}(1 + 4\sqrt{n})^2.$$

Thus, we have

$$(2.8) \quad \mathbb{P}(E) \geq \mathbb{P}(2\langle \mathbf{f}_2 - \mathbf{f}_1, \xi \rangle_2 \geq 2\sqrt{2}|\mathbf{f}_2 - \mathbf{f}_1|_2) = \mathbb{P}(Z \geq \sqrt{2}) \geq 0.07,$$

where  $Z \sim \mathcal{N}(0, 1)$ . In view of (1.2), on the event  $E$ , we have

$$\lambda_1 \leq \lambda_2 e^{-2/\beta\sqrt{n}} \leq \frac{1}{8\sqrt{n}} \leq \frac{1}{2}$$

for low temperature  $\beta$  chosen as in (2.4). Together with (2.6), it yields

$$|\mathbf{f}_\lambda|_2^2 - |\mathbf{f}_1|_2^2 \geq \frac{\sqrt{n}}{4}.$$

We now turn to the case of potentially high temperatures. Actually, the following proof holds for *any* temperature  $\beta$  as long as the  $\alpha_j$ s are chosen small enough. In this case, we can expect the  $M$  exponential weights to take comparable values. To that end, define for each  $j = 2, \dots, M$ , the event

$$F_j = \{\widehat{\text{MSE}}(f_j) \leq \widehat{\text{MSE}}(f_1)\}.$$

Define  $F = \bigcap_{j=2}^M F_j$ , and denote by  $F_j^c$  the complement of  $F_j$ . Recall that  $|\mathbf{f}_j|_2^2 = |\mathbf{f}_2|_2^2 + \alpha_j^2$  so that

$$\begin{aligned} F_j^c &= \{2\langle \mathbf{f}_j - \mathbf{f}_1, \xi \rangle_2 \leq |\mathbf{f}_j|_2^2 - |\mathbf{f}_1|_2^2\} \\ &= \{2\langle \mathbf{f}_2 - \mathbf{f}_1, \xi \rangle_2 + 2\langle \mathbf{f}_j - \mathbf{f}_2, \xi \rangle_2 \leq |\mathbf{f}_2|_2^2 - |\mathbf{f}_1|_2^2 + \alpha_j^2\} \\ &\subset E^c \cup G_j, \end{aligned}$$

where the  $E$  is defined in (2.7), and  $G_j$  is defined as

$$G_j = \{2\langle \mathbf{f}_j - \mathbf{f}_2, \xi \rangle_2 \leq \alpha_j^2 - 2\sqrt{n}\}.$$

In view of (2.5), we have

$$\mathbb{P}(G_j) \leq \mathbb{P}(2\langle \mathbf{f}_j - \mathbf{f}_2, \xi \rangle_2 \leq -\alpha_j^2) \leq \mathbb{P}(Z \geq \sqrt{2\log(100M)}) \leq \frac{0.01}{M}.$$

Therefore,

$$\mathbb{P}(F^c) \leq \mathbb{P}(E^c) + \sum_{j=2}^M \mathbb{P}(G_j) \leq 0.93 + 0.01 = 0.94.$$

Note now that on the event  $F$ , for any  $j = 2, \dots, M$ , we have  $\lambda_j \geq \lambda_1$  so that  $\lambda_1 \leq 1/M \leq 1/2$ . Together with (2.6), it yields

$$|\mathbf{f}_\lambda|_2^2 - |\mathbf{f}_1|_2^2 \geq \frac{\sqrt{n}}{2} - \frac{2n}{M} \geq \frac{\sqrt{n}}{4},$$

where, in the last inequality, we used the fact that  $M \geq 8\sqrt{n}$ .  $\square$



**2.2. Aggregate by projection.** Our lower bound for the aggregate by projection relies on a different construction of the dictionary. Let  $m$  be the smallest integer that satisfies  $m^2 \geq 4n/13$  and let  $n, M$  be large enough to ensure that  $m \geq 16$ ,  $M - 1 \geq 2m$ . Let  $\mathbf{e}^{(1)}, \dots, \mathbf{e}^{(m)} \in \mathbb{R}^n$  be the first  $m$  vectors of the canonical basis of  $\mathbb{R}^n$ . For any  $j = 1, \dots, M$ , the  $\mathbf{f}_j$ s are defined as

$$\mathbf{f}_j = \begin{cases} \sqrt{n}\mathbf{e}^{(j)}, & \text{if } 1 \leq j \leq m, \\ -\sqrt{n}\mathbf{e}^{(j)}, & \text{if } m+1 \leq j \leq 2m, \\ 0, & \text{if } j = 2m+1, \\ \mathbf{f}_1, & \text{if } j > 2m+1. \end{cases}$$

Moreover, define  $\eta \equiv 0$  so that  $0 = \text{MSE}(\mathbf{f}_{2m+1}) \leq \text{MSE}(\mathbf{f}_j)$  for all  $j \leq M$ .

**PROPOSITION 2.2.** *Fix  $n \geq 416, M \geq \sqrt{n}$ , and assume that the noise random variables  $\xi_1, \dots, \xi_n$  are i.i.d.  $\mathcal{N}(0, \sigma^2)$ . Let  $\eta$  and  $\mathcal{H}$  be defined as above. Then the projection aggregate estimator  $\mathbf{f}_{\lambda^{\text{PROJ}}}$  with weights  $\lambda^{\text{PROJ}}$  defined in (2.3) is such that*

$$\text{MSE}(\mathbf{f}_{\lambda^{\text{PROJ}}}) \geq \min_{j=1, \dots, M} \text{MSE}(\mathbf{f}_j) + \frac{\sigma^2}{\sqrt{48n}},$$

with probability larger than  $1/4$ . Moreover, the above lower bound holds with arbitrary large probability if  $n$  is chosen large enough.

**PROOF.** Note first that by homogeneity, one may assume that  $\sigma = 1$ . Next, observe that  $\mathbf{f}_{\lambda^{\text{PROJ}}} = (\mathbf{P}_m \xi, 0, \dots, 0)^\top \in \mathbb{R}^n$ , where  $\mathbf{P}_m \xi \in \mathbb{R}^m$  is the projection of  $\tilde{\xi} = (\xi_1, \dots, \xi_m)^\top$  onto  $\mathcal{B}_1^m(\sqrt{n})$ , the  $\ell_1$ -ball of  $\mathbb{R}^m$  with radius  $\sqrt{n}$ .

Let  $E$  denote the event on which  $|\tilde{\xi}|_1 \leq \sqrt{n}$  and observe that, on this event, we have  $\mathbf{P}_m \xi = \tilde{\xi}$ . It yields

$$n \text{MSE}(\mathbf{f}_{\lambda^{\text{PROJ}}}) = \sum_{j=1}^m \xi_j^2 = |\tilde{\xi}|_2^2.$$

Let now  $F$  denote the event on which  $|\tilde{\xi}|_2^2 \geq m/2$ , and note that on  $E \cap F$ , it holds

$$\text{MSE}(\mathbf{f}_{\lambda^{\text{PROJ}}}) \geq \frac{m}{2n} \geq \sqrt{\frac{1}{13n}}.$$

To conclude our proof, it remains to bound from below the probability of  $E \cap F$ . The bounds below follow from the fact that  $|\tilde{\xi}|_2^2$  follows a chi-squared distribution with  $m$  degrees of freedom. We begin by the event  $E$ . Using Hölder's inequality, we have

$$\mathbb{P}(E^c) \leq \mathbb{P}\left(|\tilde{\xi}|_2^2 \geq \frac{n}{m}\right) = \mathbb{P}\left(|\tilde{\xi}|_2^2 - \mathbb{E}|\tilde{\xi}|_2^2 \geq \frac{n}{m} - m\right).$$

Next, using the fact that  $m^2 \leq 8n/13$  together with [Laurent and Massart \[\(2000\), Lemma 1\]](#), we get

$$\mathbb{P}(E^c) \leq \mathbb{P}\left(|\tilde{\xi}|_2^2 - \mathbb{E}|\tilde{\xi}|_2^2 \geq \frac{5m}{8}\right) \leq e^{-m/16}.$$

Moreover, using [Laurent and Massart \[\(2000\), Lemma 1\]](#), we also get that

$$\mathbb{P}(F^c) = \mathbb{P}\left(|\tilde{\xi}|_2^2 - \mathbb{E}|\tilde{\xi}|_2^2 \leq -\frac{m}{2}\right) \leq e^{-m/16}.$$

Therefore, since  $n \geq 416$  implies  $m \geq 16$ , we get

$$\mathbb{P}(E \cap F) \geq 1 - \mathbb{P}(E^c) - \mathbb{P}(F^c) \geq 1 - 2e^{-m/16} \geq 1 - 2/e \geq 1/4. \quad \square$$

Note that we employed a different dictionary for each of the aggregates. Therefore, it may be the case that choosing the right aggregate for the right dictionary gives the correct deviation bounds. In the next section, we propose a new aggregate estimator called  $Q$ -aggregate, that automatically adjusts the aggregate to the dictionary at hand.

**3. Deviation optimal model selection by  $Q$ -aggregation.** According to (2.1), the weight vector  $\lambda^{\text{EXP}}$  considered in the previous section minimizes a penalized linear interpolation of the function  $\lambda \rightarrow \widehat{\text{MSE}}(\mathbf{f}_\lambda)$ . The major novelty of the method introduced in Rigollet (2012) compared to exponential weighting is to add a quadratic term to this linear interpolation. We introduce a family of estimators that extends the original estimator of Rigollet in two directions: (i) it allows for a prior weighting of the functions in the dictionary, and (ii) it allows to put different weight of each of the component of the fitting criterion via the tuning parameter  $\nu$  introduced below.

Let  $\pi \in \Lambda$  be a given prior, and define the following entropic penalty:

$$\mathcal{K}_\rho(\lambda, \pi) = \sum_{j=1}^M \lambda_j \log\left(\frac{\rho(\lambda_j)}{\pi_j}\right),$$

where  $\rho$  is a real valued function on  $[0, 1]$  that satisfies  $\rho(t) \geq t$  such that  $t \mapsto t \log \rho(t)$  is convex. We are particularly interested in the choices  $\rho = \mathbf{1}$ , the constant function equal to 1, which leads to a penalty that is linear in  $\Lambda$ , and  $\rho(t) = t$ , the identity function of  $[0, 1]$ , which leads to the well-known Kullback–Leibler penalty employed in exponential weights.

Given a dictionary  $\mathcal{H}$  and observations  $Y_1, \dots, Y_n$ , let  $Q: \Lambda \rightarrow \mathbb{R}$  be the function defined by

$$(3.1) \quad Q(\lambda) = (1 - \nu) \widehat{\text{MSE}}(\mathbf{f}_\lambda) + \nu \sum_{j=1}^M \lambda_j \widehat{\text{MSE}}(f_j) + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi),$$

where  $\nu \in [0, 1]$ . Let  $\tilde{\lambda} \in \Lambda$  be such that

$$(3.2) \quad \tilde{\lambda} \in \underset{\lambda \in \Lambda}{\operatorname{argmin}} Q(\lambda).$$

We call  $f_{\tilde{\lambda}}$  the  $Q$ -aggregate estimator. Note that on the one hand, if  $\nu = 1$  and  $\rho(t) = t$ , then  $\tilde{\lambda} = \lambda^{\text{EXP}}$ , the exponential weights defined in (1.2). On the other hand, choosing  $\nu = 0$ ,  $\rho(t) = 1$  and  $\pi$  to be the uniform prior yields  $\tilde{\lambda} = \lambda^{\text{PROJ}}$ , the projection weights.

The next theorem shows that the  $Q$ -aggregate estimator is optimal both in expectation and in deviation. It holds under less restrictive conditions on the noise random variable  $\xi_1, \dots, \xi_n$ . We say that the random vector  $\xi = (\xi_1, \dots, \xi_n)^\top$  is sub-Gaussian with variance proxy  $\sigma^2 > 0$ , if for all  $t \in \mathbb{R}^n$ , its moment generating function satisfies

$$\mathbb{E}[e^{t^\top \xi}] \leq e^{(\sigma^2 |t|_2^2)/2}.$$

Note that if  $\xi \sim \mathcal{N}_n(0, \Sigma)$ , then  $\xi$  is sub-Gaussian with variance proxy given by  $\sigma^2 = \|\Sigma\|_{\text{op}}$ , where  $\|\Sigma\|_{\text{op}}$  denotes the largest eigenvalue of the covariance matrix  $\Sigma$ .

Let  $P$  be defined on the simplex  $\Lambda$  by

$$P(\lambda) = (1 - \nu) \text{MSE}(f_\lambda) + \nu \sum_{j=1}^M \lambda_j \text{MSE}(f_j).$$

**THEOREM 3.1.** *Fix  $\nu \in (0, 1)$  and  $\pi \in \Lambda$ . Moreover, assume that the noise random variables  $\xi_1, \dots, \xi_n$  are independent and sub-Gaussian with variance proxy  $\sigma^2$ . Then for any  $\beta \geq \frac{2\sigma^2}{\min(\nu, 1-\nu)}$  and any  $\delta \in (0, 1)$ , the  $Q$ -aggregate estimator  $f_{\tilde{\lambda}}$  satisfies*

$$\text{MSE}(f_{\tilde{\lambda}}) \leq \min_{\lambda \in \Lambda} \left\{ P(\lambda) + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) + \frac{\beta}{n} \log(1/\delta) \right\},$$

with probability  $1 - \delta$ . Moreover,

$$\mathbb{E} \text{MSE}(f_{\tilde{\lambda}}) \leq \min_{\lambda \in \Lambda} \left\{ P(\lambda) + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) \right\}.$$

Theorem 3.1 follows directly from Theorem 4.1 below, so we prove only the latter in Appendix A.1.

Our theorem implies that the  $Q$ -aggregate can compete with an arbitrary  $f_\lambda$  in the convex hull with  $\lambda \in \Lambda$ . However, we are mainly interested in MS aggregation, where  $\lambda$  is at a vertex of the simplex  $\Lambda$ . With  $\nu \in (0, 1)$ , the theorem implies that the  $Q$ -aggregate estimator is deviation optimal, unlike the aggregate with exponential weights. This is explicitly stated in the following corollary, which shows that our estimator solves optimally the problem of MS aggregation. Its proof follows by simply restricting the infimum over  $\Lambda$  to the minimum over its vertices in Theorem 3.1. Nonetheless,

it is worth pointing out that our analysis focuses on deviation bounds, and it does not allow us to recover (2.2) for the aggregate with exponential weights when  $\nu = 1$ .

**COROLLARY 3.1.** *Under the assumptions of Theorem 3.1, the  $Q$ -aggregate estimator  $\mathbf{f}_{\tilde{\lambda}}$  satisfies*

$$\text{MSE}(\mathbf{f}_{\tilde{\lambda}}) \leq \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log \left( \frac{\rho(1)}{\pi_j \delta} \right) \right\},$$

with probability  $1 - \delta$ . Moreover,

$$\mathbb{E} \text{MSE}(\mathbf{f}_{\tilde{\lambda}}) \leq \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log \left( \frac{\rho(1)}{\pi_j} \right) \right\}.$$

**REMARK 3.1.** If we set  $\rho(t) = 1$  and employ the uniform prior  $\pi_j = 1/M, j = 1, \dots, M$ , then the optimization of the criterion  $Q$  is independent of  $\beta$ . In this case, we may simply set  $\nu = 1/2$ , and the  $Q$ -aggregate estimator becomes parameter free, and we recover the original aggregate of Rigollet (2012).

**4. Algorithms.** In the previous section, we introduced and analyzed the  $Q$ -aggregate estimator. It can be easily seen that if  $M$  is moderate, then it can be computed efficiently since it requires solving the convex optimization problem (3.2). The purpose of this section is to propose *greedy model averaging* (GMA) procedures that can approximately solve the  $Q$ -aggregation formulation (3.2). Moreover, GMA leads to sparse estimators (i.e., the resulting estimators only aggregate a small number of dictionary functions) that achieve the optimal deviation bounds. These algorithms are thus appealing for their simplicity and statistical interpretability.

**4.1. Approximate  $Q$ -aggregation.** Most numerical optimization algorithms do not find the exact minimum of the objective function  $Q$ , but only approximate solutions. We introduce two algorithms that minimize  $Q$  approximately, with a very specific error term for the optimization task. It relies on the following quantity. Given a dictionary  $\mathcal{H}$ , for any  $\lambda \in \Lambda$ , let  $V(\lambda)$  denote its *variance on  $\mathcal{H}$*  and be defined by

$$V(\lambda) = \sum_{j=1}^M \lambda_j \|f_j - \mathbf{f}_{\lambda}\|^2.$$

For given  $\varepsilon_V, \varepsilon > 0$ , we call  $\mathbf{f}_{\tilde{\lambda}_\varepsilon}$  an  $(\varepsilon_V, \varepsilon)$ -approximate  $Q$ -aggregate if the vector of weights  $\tilde{\lambda}_\varepsilon \in \Lambda$  is such that

$$(4.1) \quad Q(\tilde{\lambda}_\varepsilon) \leq \min_{\lambda \in \Lambda} \{Q(\lambda) + \varepsilon_V V(\lambda) + \varepsilon\}.$$

Before going into the detailed description of the algorithms we state a generalization of Theorem 3.1 that is valid not only for exact minimizers of  $Q$  but also for approximate minimizers. Hereafter, we use the convention  $0/0 = 0$ .

**THEOREM 4.1.** *Let  $\varepsilon, \varepsilon_V, \nu > 0$  be such that  $\nu + \varepsilon_V < 1$  and fix  $\pi \in \Lambda$ . Moreover, assume that the noise random variables  $\xi_1, \dots, \xi_n$  are independent sub-Gaussians with variance proxy  $\sigma^2$ . Fix any  $\theta \in (\varepsilon_V/(\nu + \varepsilon_V), 1]$ , and choose  $\beta > 0$  such that*

$$(4.2) \quad \beta \geq 2\sigma^2 \max \left\{ \frac{1}{\nu - \varepsilon_V(1 - \theta)/\theta}; \frac{1}{(1 - \theta)(1 - \nu - \varepsilon_V)} \right\}.$$

*Then for any  $\delta \in (0, 1)$ , any  $(\varepsilon_V, \varepsilon)$ -approximate  $Q$ -aggregate estimator  $\mathbf{f}_{\tilde{\lambda}_\varepsilon}$  satisfies*

$$(4.3) \quad \text{MSE}(\mathbf{f}_{\tilde{\lambda}_\varepsilon}) \leq \min_{\lambda \in \Lambda} \left\{ P(\lambda) + \varepsilon_V V(\lambda) + \frac{\varepsilon}{\theta} + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) \right\} + \frac{\beta}{n} \log(1/\delta),$$

*with probability  $1 - \delta$ . Moreover,*

$$(4.4) \quad \mathbb{E} \text{MSE}(\mathbf{f}_{\tilde{\lambda}_\varepsilon}) \leq \min_{\lambda \in \Lambda} \left\{ P(\lambda) + \varepsilon_V V(\lambda) + \frac{\varepsilon}{\theta} + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) \right\}.$$

**REMARK 4.1.** If  $\varepsilon_V = 0$ , then (4.2) reduces to  $\beta \geq 2\sigma^2 / \min(\nu; (1 - \theta)(1 - \nu))$ . Thus if  $\nu < 1/2$ , we can take  $\theta = 1 - \nu(1 - \nu)$  and  $\beta \geq 2\sigma^2/\nu$ . If  $\nu \geq 1/2$ , then for any  $\theta \in (0, 1]$ , we have  $\min(\nu; (1 - \theta)(1 - \nu)) = (1 - \theta)(1 - \nu)$ . Furthermore, if  $\varepsilon = 0$ , then in the case  $\nu \geq 1/2$  we can let  $\theta \rightarrow 0$  and obtain Theorem 3.1.

**REMARK 4.2.** Theorem 4.1 is related to PAC-Bayes-type inequalities that also employ entropy regularization. In particular, the proof involves an interpolated risk with variance correction, and such techniques have also appeared in earlier papers such as Audibert (2004) under different context.

Clearly the smaller the  $\varepsilon_V$  and  $\varepsilon$ , the better the oracle inequality. Nevertheless, in the canonical example where  $\pi$  is the uniform prior, it is sufficient to have  $\varepsilon$  uniformly bounded by  $C(\log M)/n$  for some  $C > 0$  in order to maintain a statistical accuracy of the same order as that of the true  $Q$ -aggregate. However, if an estimator has error term with  $\varepsilon = 0$  and a constant  $\varepsilon_V > 0$ , then it achieves a statistical accuracy of the same order as that of the true  $Q$ -aggregate because the variance term  $\varepsilon_V V$  vanishes at the vertices of the simplex  $\Lambda$ . This is the main reason to differentiate  $\varepsilon_V$  and  $\varepsilon$  in (4.1). As we will show later on, specially designed greedy algorithms can lead to an error term with  $\varepsilon = 0$ , and thus such greedy algorithms achieve optimal deviation bounds for MS aggregation.

**4.2. Greedy  $Q$ -aggregation.** Optimizing convex functions over convex sets is the bread and butter of modern statistical computing, with many algorithms ranging from gradient descent to interior point (IP) methods [see,

---

**Algorithm 1** Greedy model averaging (GMA-1 and GMA-1<sub>+</sub>)

---

**Input:** Noisy observation  $\mathbf{Y}$ , dictionary  $\mathcal{H} = \{f_1, \dots, f_M\}$ , prior  $\pi \in \Lambda$ , parameters  $\nu, \beta$ .

**Output:** Aggregate estimator  $\mathbf{f}_{\lambda^{(k)}}$ .

Let  $\lambda^{(0)} = 0$ ,  $\mathbf{f}_{\lambda^{(0)}} = 0$ .

**for**  $k = 1, 2, \dots$  **do**

Set  $\alpha_k = \frac{2}{k+1}$

$J^{(k)} = \arg \min_j (\nabla Q(\lambda^{(k)}))_j$

option-1 (GMA-1)  $\lambda^{(k)} = \lambda^{(k-1)} + \alpha_k (\mathbf{e}^{(J^{(k)})} - \lambda^{(k-1)})$

option-2 (GMA-1<sub>+</sub>)  $\lambda^{(k)} = \arg \min_{\lambda \in \Lambda} Q(\lambda)$  s.t.  $\lambda_j = 0$  for  $j \notin \{J^{(1)}, \dots, J^{(k)}\}$

**end for**

---

e.g., Boyd and Vandenberghe (2004) for a recent overview]. For simple constraints sets such as the simplex  $\Lambda$  considered here, so-called *proximal methods* [see, e.g., Beck and Teboulle (2009)] have shown very promising performance, especially when  $M$  becomes large. However, the most efficient of these methods (IP and proximal methods) does not output a sparse solution in a general case.

In the sequel, we focus on simple greedy model averaging algorithms (i.e., each iteration takes the form of a greedy selection of a function in the dictionary) that enjoy the following property. After  $k$  iterations, these algorithms return a vector  $\lambda^{(k)}$  such that (i),  $\lambda^{(k)}$  has at most  $k$  nonzero coefficients, and (ii)  $\mathbf{f}_{\lambda^{(k)}}$  is an approximate  $Q$ -aggregate estimator, where the quality of the approximation will be made explicit. Specifically, appropriately designed greedy algorithms can give  $\varepsilon = 0$  in (4.1) for all  $k \geq 2$ , and thus achieve optimal deviation bounds using only  $k \geq 2$  dictionary functions.

Minimizing a quadratic objective over the simplex  $\Lambda$  is a common problem in statistics and optimization. We focus on greedy algorithms introduced into the statistical literature by Jones (1992). In optimization, greedy algorithms over simplex  $\Lambda$  are known as *Frank–Wolfe*-type (or reduced gradient) methods. Their name refers to the original paper of Frank and Wolfe (1956).

We consider a few variants of greedy algorithms described in Algorithms 1 and 2. In these algorithms,  $\mathbf{e}^{(j)}$  denotes the  $j$ th vector of the canonical basis of  $\mathbb{R}^M$ . Both algorithms can be seen as greedy algorithms that add at most one function from the dictionary at each iteration. This feature is attractive as it outputs a  $k$ -sparse solution that depends on at most  $k$  functions from the dictionary after  $k$  iterations. Each algorithm contains two variants: GMA-0 and GMA-0<sub>+</sub> in Algorithm 2, and GMA-1 and GMA-1<sub>+</sub> in Algorithm 1. At the same sparsity level  $k$ , the GMA-0<sub>+</sub> (resp., GMA-1<sub>+</sub>) variant can further reduce approximation error of GMA-0 (resp., GMA-1) in (4.1) via a more aggressive optimization step. This kind of additional

---

**Algorithm 2** Greedy model averaging (GMA-0 and GMA-0<sub>+</sub>)

---

**Input:** Noisy observation  $\mathbf{Y}$ , dictionary  $\mathcal{H} = \{f_1, \dots, f_M\}$ , prior  $\pi \in \Lambda$ , parameters  $\nu, \beta$ .**Output:** Aggregate estimator  $\mathbf{f}_{\lambda^{(k)}}$ .Let  $\lambda^{(0)} = 0$ ,  $\mathbf{f}_{\lambda^{(0)}} = 0$ .**for**  $k = 1, 2, \dots$  **do**Set  $\alpha_k = \frac{2}{k+1}$  $J^{(k)} = \arg \min_j Q(\lambda^{(k-1)} + \alpha_k(\mathbf{e}^{(j)} - \lambda^{(k-1)}))$ option-1 (GMA-0)  $\lambda^{(k)} = \lambda^{(k-1)} + \alpha_k(\mathbf{e}^{(J^{(k)})} - \lambda^{(k-1)})$ option-2 (GMA-0<sub>+</sub>)  $\lambda^{(k)} = \arg \min_{\lambda \in \Lambda} Q(\lambda)$  s.t.  $\lambda_j = 0$  for  $j \notin \{J^{(1)}, \dots, J^{(k)}\}$ **end for**

---

optimization is referred to as *fully-corrective* step [Shalev-Shwartz, Srebro and Zhang (2010)], which is known to improve performance in practice. The difference between Algorithms 1 and 2 is that the former uses first order information, namely the gradient  $\nabla Q$ , to pick the best coordinate  $J^{(k)}$  (which is the standard Frank–Wolfe procedure in the greedy algorithm literature), while the latter uses only zero order information, namely, the coordinate that minimizes the objective value  $Q(\cdot)$  (which is relatively uncommon in the greedy algorithm literature). A similar algorithm with the purpose of solving MS aggregation has appeared in Dai and Zhang (2011).

Note that both algorithms give approximate solutions  $\lambda^{(k)}$  that converges to the optimal solution of (3.2); that is, when  $k \rightarrow \infty$ , we have  $\varepsilon_V \rightarrow 0$  and  $\varepsilon \rightarrow 0$  in (4.1). The classical Frank–Wolfe style analysis of greedy algorithms leads to the same convergence rate for both approaches with error term of  $\varepsilon_V = 0$  and  $\varepsilon > 0$  in (4.1). The result is presented below in Proposition 4.1. Moreover, we present a new analysis that differentiates these two algorithms. Specifically we obtain a convergence result in Theorem 4.2 below with error term of  $\varepsilon = 0$  in (4.1) for Algorithm 2 when  $k \geq 2$  (but we are unable to prove the same result for Algorithm 1). The importance of achieving error with  $\varepsilon = 0$  is that for  $k \geq 2$ , Algorithm 2 can produce a  $k$ -sparse approximate solution  $\lambda^{(k)}$  of (3.2) that achieves optimal deviation.

The following proposition follows from the standard analysis in Frank and Wolfe (1956), Jones (1992). It shows that the estimators  $\lambda^{(k)}$  from Algorithms 1 and 2 converge to the optimal solution of the  $Q$ -aggregation formulation (3.2). Therefore when  $k \rightarrow \infty$ ,  $\lambda^{(k)}$  achieves optimal deviation bound. However, a disadvantage of the bound is that the result does not imply optimal deviation bounds for  $\lambda^{(k)}$  when  $k$  is small (e.g., when  $k = 2$ ).

**PROPOSITION 4.1.** *Assume that the dictionary  $\mathcal{H}$  is such that  $\max_j \|f_j\| \leq L$ . Fix  $\nu \in (0, 1/2)$  and  $\pi \in \Lambda$ . Moreover, assume that the noise random vari-*

ables  $\xi_1, \dots, \xi_n$  are independent and sub-Gaussian with variance proxy  $\sigma^2$ . Take  $\rho = \mathbf{1}$  and

$$\beta \geq \frac{2\sigma^2}{\nu}.$$

Then, for any  $k \geq 1$ , the aggregate estimator  $\mathbf{f}_{\lambda^{(k)}}$  where  $\lambda^{(k)}$  is output by GMA-1 or GMA-0 (or GMA-1<sub>+</sub> or GMA-0<sub>+</sub>) after  $k$  steps, satisfies

$$\text{MSE}(\mathbf{f}_{\lambda^{(k)}}) \leq \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log \left( \frac{1}{\pi_j \delta} \right) \right\} + \frac{16(1-\nu)^2 L^2}{1-2\nu} \frac{1}{k+3},$$

with probability  $1 - \delta$ . Moreover,

$$\mathbb{E} \text{MSE}(\mathbf{f}_{\lambda^{(k)}}) \leq \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log \left( \frac{1}{\pi_j} \right) \right\} + \frac{16(1-\nu)^2 L^2}{1-2\nu} \frac{1}{k+3}.$$

REMARK 4.3. For simplicity, we consider the case of  $\nu < 1/2$ , although similar bounds can be obtained with  $\nu \geq 1/2$ .

REMARK 4.4. The result of Proposition 4.1 follows from the classical greedy algorithm analysis in Frank and Wolfe (1956), Jones (1992), Barron (1993). In particular, the result for  $\lambda^{(k)}$  output by GMA-1 is well known in the literature; see also Clarkson (2008), Jaggi (2011). For completeness, we include the proof in Appendix A.3 especially since the greedy step in GMA-0 (and GMA-0<sub>+</sub>) is relatively uncommon.

REMARK 4.5. It is known that the fully-corrective variants GMA-0<sub>+</sub> and GMA-1<sub>+</sub> generally achieve better performance than their partially-corrective counterparts GMA-0 and GMA-1 at the same sparsity level  $k$ . Although our analysis does not show their advantages, faster convergence rates can be obtained for fully-corrective algorithms under additional assumptions [Shalev-Shwartz, Srebro and Zhang (2010)]. Since the issue is not essential for our paper, we only illustrate the benefit of fully-corrective updates by experiments.

REMARK 4.6. It follows from the proof of Proposition 4.1 that GMA-0 can be used to optimize the function  $Q$  over the simplex  $\Lambda$ . Therefore, we can use it as a subroutine for option-2 in the description of Algorithms 2 and 1. More precisely, the following bound holds:

$$Q(\lambda^{(k)}) \leq \min_{\lambda \in \Lambda} Q(\lambda) + \frac{16(1-\nu)L^2}{k+3}.$$

For the approximation error  $\frac{16(1-\nu)^2 L^2}{1-2\nu} \frac{1}{k+3}$  to be of the same order as the estimation error, one may choose  $k$  such that

$$k \geq \frac{16(1-\nu)^2 L^2 n}{\beta(1-2\nu) \log(1/\pi_{\max})} - 3,$$



where  $\pi_{\max} = \max_j \pi_j$ . In particular, if  $\pi$  is the uniform prior, then  $\mathbf{f}_{\lambda^{(k)}}$  solves the problem of MS aggregation optimally after

$$k \geq \frac{16(1-\nu)^2 L^2 n}{\beta(1-2\nu) \log(M)} - 3$$

iterations.

Note that the above theorem requires the somewhat unpleasant assumption that the functions in the dictionary are uniformly bounded in  $\|\cdot\|$  norm. Indeed, this assumption has not appeared so far and is therefore not natural in this problem.

More importantly, the bound only leads to optimal deviation for large  $k$  of the order  $n/\log(M)$ . The cause of this unpleasant issue is that the error term is with  $\varepsilon \neq 0$  and  $\varepsilon_V = 0$  in (4.1). In order to obtain optimal deviation bound, we have to derive an error bound of the form (4.1) with either  $\varepsilon = O(\log(M)/n)$ , or with  $\varepsilon = 0$  and  $\varepsilon_V \neq 0$ . In the later case, we allow  $\varepsilon_V$  to be relatively large, which means that we do not have to solve (3.2) accurately. The following theorem shows that such an error bound (with  $\varepsilon = 0$ ) can be achieved via GMA-0 (and GMA-0<sub>+</sub>); in addition, this result removes the assumption on the boundedness of dictionary function.

**THEOREM 4.2.** *Fix  $\nu \in (0, 1)$ ,  $k \geq 2$  and  $\pi \in \Lambda$ . Moreover, assume that the noise random variables  $\xi_1, \dots, \xi_n$  are independent and sub-Gaussian with variance proxy  $\sigma^2$ . Take  $\rho = \mathbf{1}$  and*

$$\beta \geq 2\sigma^2 \inf_{\theta \in (0, 1]} \max \left\{ \frac{1}{\nu - (4(1-\nu)(1-\theta))/((k+3)\theta)}; \frac{1}{(1-\theta)(1-\nu)(1-4/(k+3))} \right\}.$$

*Then the aggregate estimator  $\mathbf{f}_{\lambda^{(k)}}$  where  $\lambda^{(k)}$  is output by GMA-0 (or GMA-0<sub>+</sub>) after  $k$  steps, satisfies*

$$\text{MSE}(\mathbf{f}_{\lambda^{(k)}}) \leq \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log \left( \frac{1}{\pi_j \delta} \right) \right\},$$

*with probability  $1 - \delta$ . Moreover,*

$$\mathbb{E} \text{MSE}(\mathbf{f}_{\lambda^{(k)}}) \leq \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n} \log \left( \frac{1}{\pi_j} \right) \right\}.$$

**REMARK 4.7.** The theorem implies deviation bounds of the optimal order for all  $k \geq 2$ , and the constant  $\beta$  decreases to  $2\sigma^2/\min(\nu, 1-\nu)$  as in Theorem 3.1 when  $k \rightarrow \infty$ . Such results indicate that the choice of  $\nu$  is not critical and any positive constant leads to the same optimal bound. However, we can optimize the constant by choosing  $\nu = 1/2$  and we use this value in the simulations.

Moreover, a careful inspection of the proof indicates that  $f_{\lambda^{(k)}}$  where  $\lambda^{(k)}$  is output by GMA-0 (or GMA-0<sub>+</sub>) after  $k$  steps is a  $(\varepsilon_V, 0)$ -approximate  $Q$ -aggregate estimator with  $\varepsilon_V = 4(1 - \nu)/(k + 3)$ . As a result, the condition  $\nu + \varepsilon_V < 1$  of Theorem 4.1 requires that  $k \geq 2$ .

To get a better quantitative idea of the result, we illustrate the particular choice  $\nu = 1/2$ . In this case, it can be easily shown that the optimal  $\theta$  is given by  $\theta_k^* = 2/(\sqrt{k+3} + 2)$ . Therefore, in this case, one may take

$$\beta \geq \frac{4\sigma^2}{1 - 2/\sqrt{k+3}}.$$

In particular, for  $k = 2$ , it is sufficient to take  $\beta = 20\sigma^2/(1 + 2/\sqrt{5}) \geq 37\sigma^2$ . Although it achieves the optimal rate for MS aggregation, the large constant implies that it is still beneficial to run the algorithm for more than two iterations. This is confirmed by our experiments.

It is worth pointing out that with flat prior, the first stage estimator  $f_{\lambda^{(1)}} = f_{\hat{j}}$  is simply the empirical risk minimizer with  $\hat{j} \in \operatorname{argmin}_j \widehat{\operatorname{MSE}}(f_j)$ . We have already pointed out that this estimator achieves sub-optimal deviation bounds; therefore the requirement of  $k \geq 2$  in our analysis is natural. With  $k = 2$ , the estimator  $f_{\lambda^{(2)}}$  is related to the STAR algorithm, which can be regarded as a two-stage greedy algorithm that minimizes the empirical loss function instead of the  $Q$ -aggregation loss investigated in this paper. This means that we cannot directly generalize the STAR algorithm to more than two stages since it converges to  $f_{\lambda^{\text{PROJ}}}$  which is known to be suboptimal for MS aggregation.

Notice that Theorem 4.2 has consequences on optimization problems beyond the scope of this paper. Indeed, we constructed a greedy algorithm for which the approximation error at each iteration is expressed as a function (here  $\varepsilon_V V$ ) and not simply a constant as usual. This construction allowed us to derive convergence rate that achieves optimal deviation bounds for greedy model averaging, and to avoid stringent and unnatural conditions on the boundedness of the problem. One of the key aspects of the function  $\varepsilon_V V$  is that it vanishes on the set of vertices. We believe that this technique may find applications in other optimization problems.

**5. Numerical experiments.** Although optimal deviation bounds are obtained for greedy  $Q$  aggregation with  $k \geq 2$ , our analysis suggests that the performance can increase when  $k$  increases (due to reduced constants). The purpose of this section is to illustrate this behavior using numerical examples. We focus on the average performance of different algorithms and configurations.

We identify a function  $f$  with a vector  $(f(x_1), \dots, f(x_n))^T \in \mathbb{R}^n$ . Define  $f_1, \dots, f_M$  so that the  $n \times M$  design matrix  $\mathbf{X} = [f_1, \dots, f_M]$  has i.i.d. standard Gaussian entries. Let  $I_n$  denote the identity matrix of  $\mathbb{R}^n$ , and let

$\Delta \sim \mathcal{N}_n(0, I_n)$  be a random vector. The regression function is defined by  $\eta = f_1 + 0.5\Delta$ . Note that typically  $f_1$  will be the closest function to  $\eta$  but not necessarily. The noise vector  $\xi \sim \mathcal{N}_n(0, 4I_n)$  is drawn independently of  $\mathbf{X}$ .

We define the oracle model (OM)  $f_{j^*}$ , where  $j^* = \operatorname{argmin}_j \operatorname{MSE}(f_j)$ . The performance difference between an estimator  $\hat{\eta}$  and the oracle model  $f_{j^*}$  is measured by the *regret* defined as

$$R(\hat{\eta}) = \operatorname{MSE}(\hat{\eta}) - \operatorname{MSE}(f_{j^*}).$$

We run GMA-0, GMA-0<sub>+</sub>, GMA-1 and GMA-1<sub>+</sub> algorithms for  $k$  iterations up to  $k = 40$ . The temperature  $\beta$  of the exponential weights (EXP) is tuned via 10-fold cross-validation. The projection aggregation (PROJ) estimator is obtained from GMA-0 with  $\nu = 0$  for 250 iterations following Remark 4.6. The fully-corrective optimization steps in GMA-0<sub>+</sub> and GMA-1<sub>+</sub> are implemented using GMA-0 and GMA-1 restricted to the support  $\{J^{(1)}, \dots, J^{(k)}\}$  at each step  $k$ . The purpose is to achieve better performance at the same sparsity level  $k$ .

Since the target is  $\eta = f_1 + 0.5\Delta$ , and  $f_1$  and  $\Delta$  are random Gaussian vectors, the oracle model is likely  $f_1$  (but it may not be  $f_1$  due to the misspecification vector  $\Delta$ ). The noise  $\sigma = 2$  is relatively large, which implies a situation where the best convex aggregation does not outperform the oracle model. This is the scenario considered in this paper. For simplicity, all algorithms use a flat prior  $\pi_j = 1/M$  for all  $j$ .

The experiment is performed with the parameters  $n = 50$ ,  $M = 200$ , and  $\sigma = 2$ , and repeated for 500 replications. In order to avoid cluttering, the detailed regret of different algorithms are given in Table 2 in the Appendix B. Table 1 is a simplified comparison of commonly used estimators (EXP and PROJ as well as STAR) with GMA-0, GMA-0<sub>+</sub>, GMA-1 and GMA-1<sub>+</sub> and  $\nu = 0.5$ . The regret is reported using the “mean  $\pm$  standard deviation” format.

TABLE 1  
*Performance comparison*

	STAR	EXP	PROJ
	0.43 $\pm$ 0.41	0.386 $\pm$ 0.47	0.407 $\pm$ 0.28

$\nu = 0.5$	$k = 1$	$k = 2$	$k = 5$	$k = 20$	$k = 40$
GMA-0	0.508 $\pm$ 0.76	0.42 $\pm$ 0.53	0.358 $\pm$ 0.42	0.336 $\pm$ 0.38	0.332 $\pm$ 0.37
GMA-0 <sub>+</sub>	0.508 $\pm$ 0.76	0.366 $\pm$ 0.5	0.341 $\pm$ 0.4	0.336 $\pm$ 0.38	0.336 $\pm$ 0.38
GMA-1	0.54 $\pm$ 0.79	0.683 $\pm$ 0.44	0.391 $\pm$ 0.38	0.342 $\pm$ 0.36	0.334 $\pm$ 0.37
GMA-1 <sub>+</sub>	0.54 $\pm$ 0.79	0.381 $\pm$ 0.46	0.338 $\pm$ 0.38	0.336 $\pm$ 0.38	0.336 $\pm$ 0.38

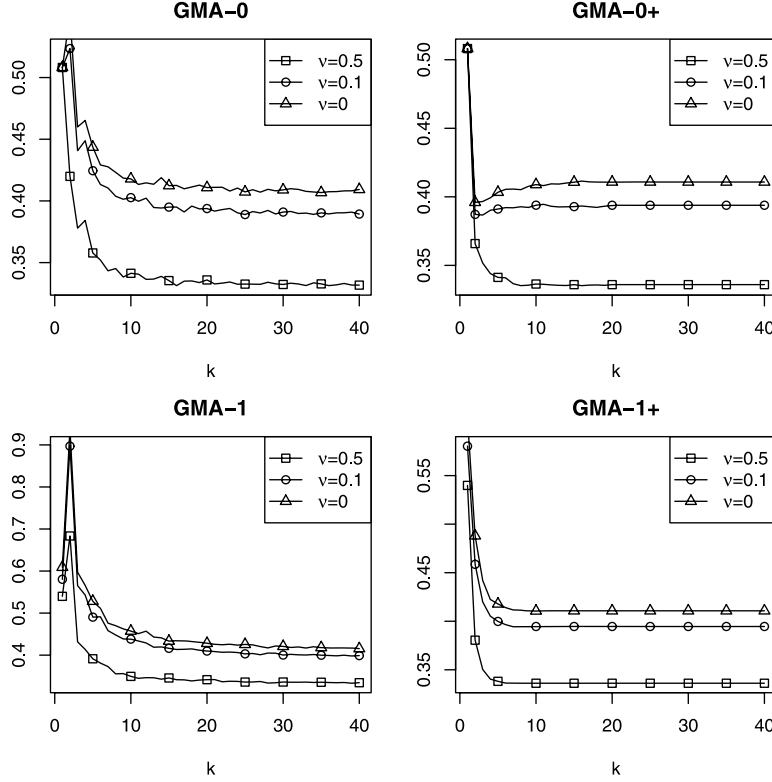


FIG. 1. Regrets  $R(f_{\lambda^{(k)}})$  versus iterations  $k$  for  $\nu = 0.5, 0.1, 0$ , under 500 replications.

The results in Table 1 indicate that for GMA-0 (or GMA-0<sub>+</sub>), from  $k = 1$  (corresponding to MS aggregation) to  $k = 2$ , there is significant reduction of error. The performance of GMA-0 (or GMA-0<sub>+</sub>) with  $k = 2$  is comparable to that of the STAR algorithm. This is not surprising as STAR can be regarded as the stage-2 greedy model averaging estimator based on empirical risk minimization. We also observe that the error keeps decreasing (but at a slower pace) when  $k > 2$ , which is consistent with Theorem 4.2. It means that in order to achieve good performance, it is necessary to use more stages than  $k = 2$  [although this does not change the  $O(1/n)$  rate for the regret, it can significantly reduce the constant]. It becomes better than EXP when  $k$  is as small as 5, which still gives a relatively sparse averaged model.

Figure 1 compares the MSE performance of different values of  $\nu$  for greedy algorithms considered in the paper. It shows that for the scenario we are interested in (i.e., where the noise is relatively large, and the best single model is nearly as good as the best convex hull combination), it is beneficial to choose  $\nu = 0.5$ . Note that the greedy procedure with  $\nu = 0$  converges to the convex hull projection aggregate estimator  $f_{\lambda^{\text{PROJ}}}$  which we have shown

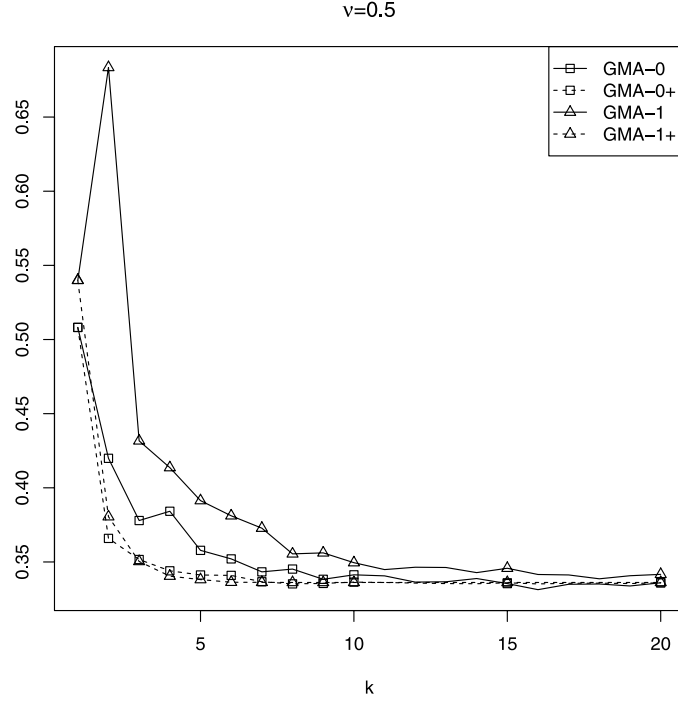


FIG. 2. Regrets  $R(f_{\lambda(k)})$  versus iterations  $k$  of different greedy procedures at  $\nu = 0.5$ , under 500 replications.

to be sub-optimal for MS aggregation. Therefore these results are consistent with our theoretical analysis, and illustrate the importance of  $Q$ -aggregation with  $\nu > 0$  for MS aggregation.

Figure 2 compares the MSE of different greedy procedures at  $\nu = 0.5$  (additional comparisons at  $\nu = 0.1$  and  $\nu = 0$  can be found in Figure 3 in the Appendix B). It shows that the classical first order greedy method GMA-1 generally performs worse than GMA-0 for all  $k$  and especially when  $k$  is small. This is consistent with our theoretical analysis since Theorem 4.2 only applies to GMA-0. The experiments show that the fully-corrective variants GMA-0<sub>+</sub> and GMA-1<sub>+</sub> can potentially give more accurate results than GMA-0 and GMA-1 at the same sparsity level  $k$ .

## APPENDIX A: PROOFS

**A.1. Proof of Theorem 4.1.** Let  $\tilde{\lambda}$  be such that

$$Q(\tilde{\lambda}) \leq \min_{\lambda \in \Lambda} \{Q(\lambda) + \varepsilon_V V(\lambda) + \varepsilon\}.$$

Fix  $\theta \in (0, 1)$  and for any  $\lambda \in \Lambda$ , define  $\lambda_\theta \in \Lambda$  by  $\lambda_\theta = (1 - \theta)\tilde{\lambda} + \theta\lambda$ .

Note that

$$P(\tilde{\lambda}) - P(\lambda_\theta) = (1 - \nu)[\text{MSE}(\mathbf{f}_{\tilde{\lambda}}) - \text{MSE}(\mathbf{f}_{\lambda_\theta})] + \nu\theta \sum_{j=1}^M (\tilde{\lambda}_j - \lambda_j) \text{MSE}(f_j).$$

Moreover, it is not hard to verify that

$$\text{MSE}(\mathbf{f}_{\tilde{\lambda}}) - \text{MSE}(\mathbf{f}_{\lambda_\theta}) = \theta \text{MSE}(\mathbf{f}_{\tilde{\lambda}}) - \theta \text{MSE}(\mathbf{f}_\lambda) + \theta(1 - \theta) \|\mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_\lambda\|^2.$$

The above two displays and the definition of  $P(\lambda)$  yield

$$(A.1) \quad P(\tilde{\lambda}) - P(\lambda_\theta) = \theta[P(\tilde{\lambda}) - P(\lambda)] + \theta(1 - \theta)(1 - \nu) \|\mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_\lambda\|^2.$$

Moreover, by the definition of  $\tilde{\lambda}$ , we have

$$Q(\tilde{\lambda}) \leq Q(\lambda_\theta) + \varepsilon_V V(\lambda_\theta) + \varepsilon.$$

By replacing  $Q(\tilde{\lambda})$  and  $Q(\lambda_\theta)$  with the expansion

$$Q(\lambda) = P(\lambda) + \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle - 2\langle \boldsymbol{\xi}, \mathbf{f}_\lambda - \eta \rangle + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi),$$

where  $\boldsymbol{\xi} = \mathbf{Y} - \eta$ , we obtain

$$\begin{aligned} P(\tilde{\lambda}) - P(\lambda_\theta) &\leq 2\langle \boldsymbol{\xi}, \mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_{\lambda_\theta} \rangle + \frac{\beta}{n} \mathcal{K}_\rho(\lambda_\theta, \pi) - \frac{\beta}{n} \mathcal{K}_\rho(\tilde{\lambda}, \pi) + \varepsilon_V V(\lambda_\theta) + \varepsilon \\ &\leq 2\langle \boldsymbol{\xi}, \mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_{\lambda_\theta} \rangle + \frac{\beta\theta}{n} \mathcal{K}_\rho(\lambda, \pi) - \frac{\beta\theta}{n} \mathcal{K}_\rho(\tilde{\lambda}, \pi) + \varepsilon_V V(\lambda_\theta) + \varepsilon, \end{aligned}$$

where in the second inequality, we applied Jensen's inequality with  $\lambda_\theta = (1 - \theta)\tilde{\lambda} + \theta\lambda$  to the convex function  $\lambda \mapsto \mathcal{K}_\rho(\lambda, \pi)$ . Plugging (A.1) into this and dividing by  $\theta$ , we get

$$(A.2) \quad \begin{aligned} P(\tilde{\lambda}) - P(\lambda) &\leq \tilde{R}_n(\mathbf{f}_\lambda) - (1 - \theta)(1 - \nu) \|\mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_\lambda\|^2 \\ &\quad + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) + \frac{\varepsilon_V}{\theta} V(\lambda_\theta) + \frac{\varepsilon}{\theta}, \end{aligned}$$

where, using the fact that  $\mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_{\lambda_\theta} = \theta(\mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_\lambda)$ , we can take

$$\tilde{R}_n(\mathbf{f}_\lambda) = 2\langle \boldsymbol{\xi}, \mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_\lambda \rangle - \frac{\beta}{n} \mathcal{K}_\rho(\tilde{\lambda}, \pi).$$

The following lemma allows us to control  $\tilde{R}_n(\mathbf{f}_\lambda)$  both in expectation and with high probability.

LEMMA A.1. *Let the noise vector  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$  be sub-Gaussian with variance proxy  $\sigma^2$ . Then, for any  $\beta > 0$ ,  $\lambda \in \mathbb{R}^M$ , we have*

$$\mathbb{E} \exp \left( \frac{n}{\beta} \tilde{R}_n(\mathbf{f}_\lambda) - \frac{2\sigma^2 n}{\beta^2} \sum_{j=1}^M \tilde{\lambda}_j \Upsilon_j(\lambda) \right) \leq 1,$$

where  $\Upsilon_j(\lambda) = \|f_j - \mathbf{f}_\lambda\|^2$ .

PROOF. Fix  $\lambda \in \mathbb{R}^M$ . Using successively Jensen's inequality and the assumption that  $t \leq \rho(t)$  yields

$$\begin{aligned}
& \mathbb{E} \exp \left( \frac{n \tilde{R}_n(\mathbf{f}_\lambda)}{\beta} - \frac{2\sigma^2 n}{\beta^2} \sum_{j=1}^M \tilde{\lambda}_j \Upsilon_j(\lambda) \right) \\
&= \mathbb{E} \exp \left[ \sum_{j=1}^M \tilde{\lambda}_j \left( \frac{2n}{\beta} \langle \boldsymbol{\xi}, f_j - \mathbf{f}_\lambda \rangle - \log \left( \frac{\rho(\tilde{\lambda}_j)}{\pi_j} \right) - \frac{2\sigma^2 n}{\beta^2} \Upsilon_j(\lambda) \right) \right] \\
&\leq \mathbb{E} \sum_{j=1}^M \tilde{\lambda}_j \exp \left( \frac{2n}{\beta} \langle \boldsymbol{\xi}, f_j - \mathbf{f}_\lambda \rangle - \log \left( \frac{\rho(\tilde{\lambda}_j)}{\pi_j} \right) - \frac{2\sigma^2 n}{\beta^2} \Upsilon_j(\lambda) \right) \\
&\leq \sum_{j=1}^M \pi_j \mathbb{E} \exp \left( \frac{2n}{\beta} \langle \boldsymbol{\xi}, f_j - \mathbf{f}_\lambda \rangle - \frac{2\sigma^2 n}{\beta^2} \Upsilon_j(\lambda) \right).
\end{aligned}$$

Observe now that since  $\boldsymbol{\xi}$  is sub-Gaussian, we have

$$\mathbb{E} \exp \left( \frac{2n}{\beta} \langle \boldsymbol{\xi}, f_j - \mathbf{f}_\lambda \rangle \right) \leq \exp \left( \frac{2n\sigma^2}{\beta^2} \|f_j - \mathbf{f}_\lambda\|^2 \right) = \exp \left( \frac{2n\sigma^2}{\beta^2} \Upsilon_j(\lambda) \right).$$

This completes the proof of our lemma.  $\square$

To prove the first result of Theorem 4.1, note that Lemma A.1 together with a Chernoff bound yield that for any  $\delta \in (0, 1)$ ,

$$(\text{A.3}) \quad \tilde{R}_n(\mathbf{f}_\lambda) \leq \frac{2\sigma^2}{\beta} \sum_{j=1}^M \tilde{\lambda}_j \|f_j - \mathbf{f}_\lambda\|^2 + \frac{\beta \log(1/\delta)}{n}$$

with probability at least  $1 - \delta$ . By combining (A.2) and (A.3), and using the definition of  $P(\tilde{\lambda})$ , we obtain

$$\begin{aligned}
& (1 - \nu) \text{MSE}(\mathbf{f}_{\tilde{\lambda}}) + \nu \sum_{j=1}^M \tilde{\lambda}_j \text{MSE}(f_j) \\
(\text{A.4}) \quad & \leq P(\lambda) + \frac{2\sigma^2}{\beta} \sum_{j=1}^M \tilde{\lambda}_j \|f_j - \mathbf{f}_\lambda\|^2 + \frac{\beta \log(1/\delta)}{n} \\
& \quad - (1 - \theta)(1 - \nu) \|\mathbf{f}_{\tilde{\lambda}} - \mathbf{f}_\lambda\|^2 + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) + \frac{\varepsilon_V}{\theta} V(\lambda_\theta) + \frac{\varepsilon}{\theta}.
\end{aligned}$$

The following identities follows directly from algebra:

$$\sum_{j=1}^M \tilde{\lambda}_j \text{MSE}(f_j) = \text{MSE}(\mathbf{f}_{\tilde{\lambda}}) + V(\tilde{\lambda}),$$

$$\sum_{j=1}^M \tilde{\lambda}_j \|f_j - f_\lambda\|^2 = V(\tilde{\lambda}) + \|f_{\tilde{\lambda}} - f_\lambda\|^2.$$

Together with (A.4), they yield

$$\begin{aligned} \text{MSE}(f_{\tilde{\lambda}}) &\leq P(\lambda) + \left( \frac{2\sigma^2}{\beta} - \nu \right) V(\tilde{\lambda}) + \frac{\beta \log(1/\delta)}{n} \\ (A.5) \quad &+ \left[ \frac{2\sigma^2}{\beta} - (1-\theta)(1-\nu) \right] \|f_{\tilde{\lambda}} - f_\lambda\|^2 + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \pi) \\ &+ \frac{\varepsilon_V}{\theta} V(\lambda_\theta) + \frac{\varepsilon}{\theta}. \end{aligned}$$

We now use the following identity which again follows directly from algebra:

$$V(\lambda_\theta) = \theta V(\lambda) + (1-\theta)V(\tilde{\lambda}) + \theta(1-\theta)\|f_{\tilde{\lambda}} - f_\lambda\|^2.$$

Together with (A.5), we obtain

$$\text{MSE}(f_{\tilde{\lambda}}) \leq P(\lambda) + G_1 V(\tilde{\lambda}) + G_2 \|f_{\tilde{\lambda}} - f_\lambda\|^2 + \frac{\beta}{n} \mathcal{K}_\rho(\lambda, \delta\pi) + \varepsilon_V V(\lambda) + \frac{\varepsilon}{\theta},$$

where  $\delta\pi = (\delta\pi_1, \dots, \delta\pi_M)^\top$ ,

$$G_1 = \frac{2\sigma^2}{\beta} - \nu + \frac{\varepsilon_V(1-\theta)}{\theta}$$

and

$$G_2 = \frac{2\sigma^2}{\beta} - (1-\theta)(1-\nu) + \varepsilon_V(1-\theta).$$

To complete the proof of (4.3), it is sufficient to note that choosing  $\beta$  as in (4.2) ensures that  $G_1 \leq 0$  and  $G_2 \leq 0$ .

Using the convexity inequality  $t \leq e^t - 1$  for any  $t \in \mathbb{R}$ , it yields that (A.3) also holds in expectation. The proof of (4.4) is then concluded in the same way as the proof of (4.3) by making statements in expectation instead of statements that hold with high probability.

**A.2. Proof of Theorem 4.2.** It follows from a Taylor expansion that for any  $\mu, \mu' \in \Lambda$ , we have

$$(A.6) \quad Q(\mu) = Q(\mu') + (\mu - \mu')^\top \nabla Q(\mu') + (1-\nu)\|f_\mu - f_{\mu'}\|^2.$$

Observe also that for any  $\lambda \in \Lambda$ , we have (both for GMA-0 and GMA-0<sub>+</sub>)

$$Q(\lambda^{(k+1)}) \leq \sum_{j=1}^M \lambda_j Q(\lambda^{(k)} + \alpha_{k+1}(\mathbf{e}^{(j)} - \lambda^{(k)})).$$



Expanding each term on the right-hand side using (A.6) with  $\mu = \lambda^{(k)} + \alpha_{k+1}(\mathbf{e}^{(j)} - \lambda^{(k)})$  and  $\mu' = \lambda^{(k)}$  yields

$$(A.7) \quad \begin{aligned} Q(\lambda^{(k+1)}) &\leq Q(\lambda^{(k)}) + \alpha_{k+1}^2(1-\nu) \sum_{j=1}^M \lambda_j \|f_j - \mathbf{f}_{\lambda^{(k)}}\|^2 \\ &\quad + \alpha_{k+1}(\lambda - \lambda^{(k)})^\top \nabla Q(\lambda^{(k)}). \end{aligned}$$

Note that

$$\sum_{j=1}^M \lambda_j \|f_j - \mathbf{f}_{\lambda^{(k)}}\|^2 = \sum_{j=1}^M \lambda_j \|f_j - \mathbf{f}_\lambda\|^2 + \|\mathbf{f}_{\lambda^{(k)}} - \mathbf{f}_\lambda\|^2.$$

Moreover, applying (A.6) with  $\mu = \lambda$  and  $\mu' = \lambda^{(k)}$  yields

$$\alpha_{k+1}(\lambda - \lambda^{(k)})^\top \nabla Q(\lambda^{(k)}) = \alpha_{k+1}[Q(\lambda) - Q(\lambda^{(k)})] - (1-\nu)\alpha_{k+1}\|\mathbf{f}_{\lambda^{(k)}} - \mathbf{f}_\lambda\|^2.$$

Plugging the above two displays into (A.7) and using  $\alpha_{k+1}^2 - \alpha_{k+1} \leq 0$ , we get

$$Q(\lambda^{(k+1)}) \leq Q(\lambda^{(k)}) + \alpha_{k+1}^2(1-\nu) \sum_{j=1}^M \lambda_j \|f_j - \mathbf{f}_\lambda\|^2 + \alpha_{k+1}[Q(\lambda) - Q(\lambda^{(k)})].$$

This can be written as

$$(A.8) \quad \delta_{k+1} \leq (1 - \alpha_{k+1})\delta_k + \alpha_{k+1}^2 B,$$

where

$$\delta_k = Q(\lambda^{(k)}) - Q(\lambda), \quad B = (1-\nu) \sum_{j=1}^M \lambda_j \|f_j - \mathbf{f}_\lambda\|^2.$$

To conclude that

$$(A.9) \quad \delta_k \leq \frac{4B}{k+3},$$

we proceed by induction on  $k$ . It is easy to see from (A.8) with  $k=0$  and  $\alpha_1=1$  that  $\delta_1 \leq B$ .

Now for  $k \geq 1$ , bound (A.8) yields

$$\begin{aligned} \delta_{k+1} &\leq \left(1 - \frac{2}{2+k}\right) \delta_k + \left(\frac{2}{2+k}\right)^2 B \\ &\leq \left(1 - \frac{2}{2+k}\right) \frac{4B}{k+3} + \left(\frac{2}{2+k}\right)^2 B = \frac{4(k^2 + 3k + 3)B}{(k+2)^2(k+3)} \leq \frac{4B}{k+4}, \end{aligned}$$

where in the second inequality, we used (A.9). We have proved that for any  $\lambda$ , it holds

$$Q(\lambda^{(k)}) \leq Q(\lambda) + \frac{4(1-\nu)}{k+3} \sum_{j=1}^M \lambda_j \|f_j - f_\lambda\|^2.$$

To complete the proof, we check that the assumptions of Theorem 4.1 with  $\varepsilon_V = 4(1-\nu)/(k+3)$  and  $\varepsilon = 0$  are satisfied. Moreover, using expression (4.2), we get the desired bound on  $\beta$ . To conclude, notice that  $V(\lambda)$  vanishes at the vertices of the simplex  $\Lambda$ .

**A.3. Proof of Proposition 4.1.** Similarly to the proof of Theorem 4.2, for both GMA-1 and GMA-1<sub>+</sub>, we have

$$\begin{aligned} & Q(\lambda^{(k+1)}) - (1-\nu)\alpha_{k+1}^2 \|f_{J^{(k)}} - f_{\lambda^{(k)}}\|^2 \\ &= Q(\lambda^{(k)}) + \alpha_{k+1}(\mathbf{e}^{(J^{(k)})} - \lambda^{(k)})^\top \nabla Q(\lambda^{(k)}) \\ &\leq \sum_{j=1}^M \lambda_j [Q(\lambda^{(k)}) + \alpha_{k+1}(\mathbf{e}^{(j)} - \lambda^{(k)})^\top \nabla Q(\lambda^{(k)})] \\ &= Q(\lambda^{(k)}) + \alpha_{k+1} \nabla Q(\lambda^{(k)})^\top (\lambda - \lambda^{(k)}) \\ &= Q(\lambda^{(k)}) + \alpha_{k+1} [Q(\lambda) - Q(\lambda^{(k)}) - (1-\nu)\|f_\lambda - f_{\lambda^{(k)}}\|^2]. \end{aligned}$$

Using  $\|f_{J^{(k)}} - f_{\lambda^{(k)}}\|^2 \leq 4L^2$ , we obtain

$$(A.10) \quad \delta_{k+1} \leq (1 - \alpha_{k+1})\delta_k + \alpha_{k+1}^2 B',$$

where we define

$$\delta_k = Q(\lambda^{(k)}) - Q(\lambda), \quad B' = 4(1-\nu)L^2.$$

Note that (A.10) also holds for GMA-0 and GMA-0<sub>+</sub> due to (A.8). Therefore similarly to the proof of Theorem 4.2, we can solve the recursion in (A.10) to obtain

$$\delta_k \leq \frac{4B'}{k+3} = \frac{16(1-\nu)L^2}{k+3}.$$

That is, we have

$$Q(\lambda^{(k)}) \leq \min_{\lambda \in \Lambda} Q(\lambda) + \frac{16(1-\nu)L^2}{k+3}.$$

We can thus apply Theorem 4.1 with  $\varepsilon_V = 0$ ,  $\varepsilon = 16(1-\nu)L^2/(k+3)$ , and  $\theta = (1-2\nu)/(1-\nu)$  to complete the proof.

## APPENDIX B: DETAILED PERFORMANCE TABLE AND FIGURES

TABLE 2  
*Regret of different algorithms: oracle model is superior to averaged models*

	STAR	EXP	PROJ
	$0.43 \pm 0.41$	$0.386 \pm 0.47$	$0.407 \pm 0.28$

	$k = 1$	$k = 2$	$k = 5$	$k = 20$	$k = 40$
GMA-0					
$\nu = 0.5$	$0.508 \pm 0.76$	$0.42 \pm 0.53$	$0.358 \pm 0.42$	$0.336 \pm 0.38$	$0.332 \pm 0.37$
$\nu = 0.1$	$0.508 \pm 0.76$	$0.523 \pm 0.5$	$0.424 \pm 0.35$	$0.394 \pm 0.3$	$0.389 \pm 0.3$
$\nu = 0$	$0.508 \pm 0.76$	$0.55 \pm 0.48$	$0.444 \pm 0.34$	$0.411 \pm 0.29$	$0.409 \pm 0.28$
GMA-0+					
$\nu = 0.5$	$0.508 \pm 0.76$	$0.366 \pm 0.5$	$0.341 \pm 0.4$	$0.336 \pm 0.38$	$0.336 \pm 0.38$
$\nu = 0.1$	$0.508 \pm 0.76$	$0.387 \pm 0.44$	$0.391 \pm 0.33$	$0.394 \pm 0.3$	$0.394 \pm 0.3$
$\nu = 0$	$0.508 \pm 0.76$	$0.396 \pm 0.43$	$0.403 \pm 0.32$	$0.411 \pm 0.29$	$0.411 \pm 0.29$
GMA-1					
$\nu = 0.5$	$0.54 \pm 0.79$	$0.683 \pm 0.44$	$0.391 \pm 0.38$	$0.342 \pm 0.36$	$0.334 \pm 0.37$
$\nu = 0.1$	$0.58 \pm 0.83$	$0.897 \pm 0.35$	$0.49 \pm 0.31$	$0.41 \pm 0.29$	$0.399 \pm 0.29$
$\nu = 0$	$0.609 \pm 0.84$	$0.937 \pm 0.32$	$0.528 \pm 0.3$	$0.428 \pm 0.27$	$0.415 \pm 0.28$
GMA-1+					
$\nu = 0.5$	$0.54 \pm 0.79$	$0.381 \pm 0.46$	$0.338 \pm 0.38$	$0.336 \pm 0.38$	$0.336 \pm 0.38$
$\nu = 0.1$	$0.58 \pm 0.83$	$0.459 \pm 0.45$	$0.4 \pm 0.31$	$0.395 \pm 0.3$	$0.395 \pm 0.3$
$\nu = 0$	$0.609 \pm 0.84$	$0.488 \pm 0.45$	$0.418 \pm 0.3$	$0.411 \pm 0.29$	$0.411 \pm 0.29$

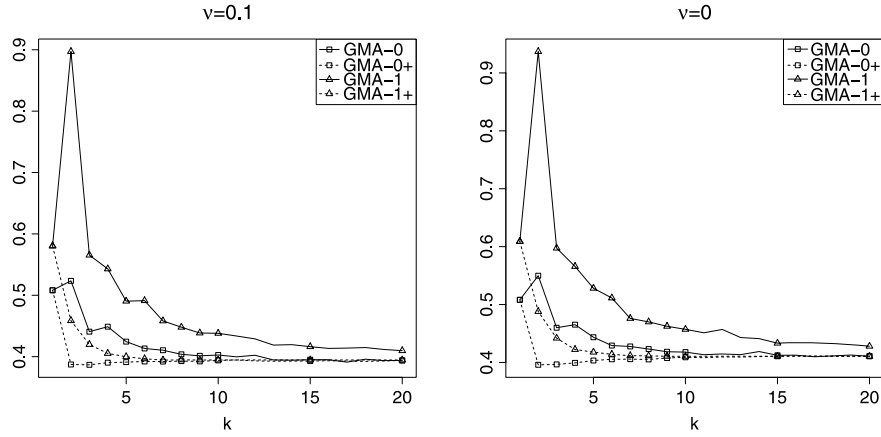


FIG. 3. Regrets  $R(f_{\lambda^{(k)}})$  versus iterations  $k$  of different greedy procedures at  $\nu = 0.1$  and  $\nu = 0$ , under 500 replications.

**Acknowledgment.** We would like to thank an anonymous referee for suggesting an improvement of bound (4.2).

## REFERENCES

- AUDIBERT, J.-Y. (2004). Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré Probab. Stat.* **40** 685–736. [MR2096215](#)
- AUDIBERT, J.-Y. (2008). Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20* (J. C. PLATT, D. KOLLER, Y. SINGER and S. ROWEIS, eds.) 41–48. MIT Press, Cambridge, MA.
- BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** 930–945. [MR1237720](#)
- BECK, A. and TEOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. [MR2486527](#)
- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](#)
- CATONI, O. (1999). Universal aggregation rules with exact bias bounds. Technical report, Laboratoire de Probabilités et Modèles Aléatoires, Paris.
- CLARKSON, K. L. (2008). Coresets, sparse greedy approximation, and the Frank–Wolfe algorithm. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 922–931. ACM, New York. [MR2487663](#)
- DAI, D. and ZHANG, T. (2011). Greedy model averaging. In *Advances in Neural Information Processing Systems 24* (J. SHAWE-TAYLOR, R. S. ZEMEL, P. BARTLETT, F. C. N. PEREIRA and K. Q. WEINBERGER, eds.) 1242–1250. MIT Press, Cambridge, MA.
- DALALYAN, A. S. and SALMON, J. (2011). Sharp oracle inequalities for aggregation of affine estimators. Available at arXiv:[1104.3969](#).
- DALALYAN, A. S. and TSYBAKOV, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory. Lecture Notes in Computer Science* **4539** 97–111. Springer, Berlin. [MR2397581](#)
- DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning* **72** 39–61.
- FRANK, M. and WOLFE, P. (1956). An algorithm for quadratic programming. *Naval Res. Logist. Quart.* **3** 95–110. [MR0089102](#)
- GAÏFFAS, S. and LECUÉ, G. (2011). Hyper-sparse optimal aggregation. *J. Mach. Learn. Res.* **12** 1813–1833. [MR2819018](#)
- JAGGI, M. (2011). Convex optimization without projection steps. Available at arXiv:[1108.1170v6](#).
- JONES, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.* **20** 608–613. [MR1150368](#)
- JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712. [MR1792783](#)
- JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. [MR2458184](#)
- LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. [MR1805785](#)
- LECUÉ, G. and MENDELSON, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145** 591–613. [MR2529440](#)

- LECUÉ, G. and MENDELSON, S. (2012). On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*. To appear.
- NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1998)*. *Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. [MR1775640](#)
- RIGOLLET, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665.
- RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. [MR2816337](#)
- RIGOLLET, P. and TSYBAKOV, A. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* To appear. Available at arXiv:[1108.5116](#).
- SHALEV-SHWARTZ, S., SREBRO, N. and ZHANG, T. (2010). Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. Optim.* **20** 2807–2832. [MR2721156](#)
- TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *COLT* (B. SCHÖLKOPF and M. K. WARMUTH, eds.). *Lecture Notes in Computer Science* **2777** 303–313. Springer, Berlin.
- YANG, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9** 475–499. [MR1707850](#)

D. DAI  
T. ZHANG  
DEPARTMENT OF STATISTICS  
RUTGERS UNIVERSITY  
PISCATAWAY, NEW JERSEY 08854  
USA  
E-MAIL: [dongdai@stat.rutgers.edu](mailto:dongdai@stat.rutgers.edu)  
[tzhang@stat.rutgers.edu](mailto:tzhang@stat.rutgers.edu)

P. RIGOLLET  
DEPARTMENT OF OPERATIONS RESEARCH  
AND FINANCIAL ENGINEERING  
PRINCETON UNIVERSITY  
PRINCETON, NEW JERSEY 08544  
USA  
E-MAIL: [rigollet@princeton.edu](mailto:rigollet@princeton.edu)